Task 2: Quantitative Trade-Off Analysis

To improve the relevance of chatbot answers, the engineering team proposed two mutually exclusive upgrades. This analysis compares both options in terms of **latency**, **cost**, and **relevance impact**, and provides a clear recommendation based on current system constraints.

Option A: Add a Cohere Re-ranker

Hypothesis:
Adding a re-ranker will improve response quality by reordering the top 10 retrieved chunks, selecting the most relevant 4 to send to the generator. This improves answer precision without increasing the input size for the LLaMA 3 model.

- Component Affected: Retrieval
- Latency Impact: +600ms per query
- Cost: $1.00 per 1,000 queries → $100/month for 100,000 queries
- Token Load Impact: None — token count remains constant

Conclusion:
Option A offers a low-risk, cost-effective method for improving answer relevance. Although it increases latency by 600ms, it avoids adding load to the generation step and controls costs.

Option B: Increase Context Size (k=4 → k=10)

Hypothesis:
By retrieving more chunks (6 additional per query), the chatbot will have richer context to generate more complete answers. However, this increases both input size and processing time for the LLaMA 3 model.

- Component Affected: Retrieval and Generation
- Retrieval Latency: +250ms per query
- Token Load Impact: 6 extra chunks × 400 tokens = 2,400 tokens per query
- Monthly Token Load: 2,400 tokens × 100,000 queries = 240,000,000 tokens

- Monthly Cost Increase: 240M ÷ 1M × $3 = **$720/month**

Conclusion:
While this option may improve recall by adding more context, it worsens the system's existing latency issues. Generation time increases significantly, and cost rises more than 7× compared to Option A.

Comparison Summary

Metric	**Option A: Re-ranker**	**Option B: k=10 Retrieval**
Latency Impact	+600ms	+250ms (retrieval) + generation delay
Monthly Cost	$100	$720
Token Load	No change	+240M tokens/month
Component Affected	Retrieval only	Retrieval and Generation
Generation Stress	None	High
Relevance Gain	Higher precision (top-4)	Higher recall (more context)
Risk of SLA Violation	Medium	High

Final Recommendation

I recommend Option A: Add a Cohere Re-ranker.

This solution provides better answer quality through smarter selection of context without adding pressure to the already strained LLaMA 3 generation step. It keeps both latency and cost within acceptable limits and aligns with the system’s current needs.

Option B may offer marginal gains in relevance, but it comes at a steep cost and increases the likelihood of SLA violations. Given that generation is already the primary performance bottleneck, Option B would make the problem worse.

Option A is the safer, more scalable choice.